Running Head: Diversity and core collection of bladder clover

**Strategies for developing a core collection of bladder clover (*Trifolium spumosum*
L.) using ecological and agro-morphological data**

K. GHAMKHAR^{A,D}, R. SNOWBALL^{A,B}, B. J. WINTLE^B, and A. H. D. BROWN^C

^ACentre for Legumes in Mediterranean Agriculture (CLIMA), University of Western Australia,
Crawley, Western Australia, 6009, Australia.

^BDepartment of Agriculture and Food Western Australia, 3 Baron-Hay Court, South Perth,
Western Australia, 6151, Australia.

^CCentre for Plant Biodiversity Research, CSIRO Plant Industry, Canberra, ACT 2601, Australia

^DCorresponding Author. Email: kioumars@cyllene.uwa.edu.au

Abstract. Core collections are a way to improve the functioning of germplasm collections through systematic evaluation, exploitation and improved management of the whole collection. The published sizes of cores have ranged from 10 to 30 percent of the collection. The first step in developing a core collection is usually stratification of the whole collection. We adopted three approaches in the process of core designation. First, the existing ecological data and newly acquired agro-morphological (evaluation) data were analysed to determine the optimal method for capturing the ecological and morphological diversity of accessions. Second, different selection strategies were compared to identify the most powerful analysis with the greatest likelihood of covering maximum diversity. Third, the correlation between the two data sets was investigated. The annual pasture legume, bladder clover (*Trifolium spumosum* L.) was used for this pilot study, as 80% of the existing 398 accessions have near complete ecological data. The study found that a two-step selection process using ecological data and stratified proportional strategy followed by combined datasets and maximizing strategy best represents the whole collection of *T. spumosum* in the core.

Additional keywords: Canonical Redundancy Analysis (RDA); Core size; Ecogeography; GIS; MSTRAT; Sampling strategy

1 Introduction

2
3 Since first advocated two decades ago, core collections have assumed a steadily increasing
4 managerial role in collections of plant genetic resources (van Hintum *et al.* 2000). A core
5 collection consists of a limited number of accessions chosen to represent the genetic spectrum
6 and include as much genetic diversity as possible of the whole collection. The overall aims of the
7 core collection are to improve both the use and conservation of the whole collection. Accessions
8 excluded from the core are held in reserve for future use, composing a back-up for the core,
9 which also acts as a reference to the whole collection (Brown 1989a).

10 Three decisions are crucial in designing a core collection. The first is the size of the core in
11 relation to the size of the whole collection. The second is the kind of data to be used. The third is
12 the procedures used to choose the core. As a guide for size, Brown (1989a, b) suggested that
13 10% of the whole collection is the optimal size of the core collection. Proportionate sizes larger
14 than 10% in theory lead only to marginally higher allele retention. In theory, a subset of 10% of
15 the original, keeps an average of > 70% of the genetic diversity of the whole collection with 95%
16 certainty. In practice, some authors have favoured smaller fractions (< 10%, Charmet and
17 Balfourier 1995; Bisht *et al.* 1998); while others have favored larger. Balfourier *et al.* (1999)
18 showed that while the standard size is a useful yardstick, it would be too small if biased and
19 highly favoured traits dominate the choice. Diwan *et al.* (1995) and Yonezawa *et al.* (1995) also
20 favoured larger fractions for the same reason. Noirot *et al.* (1996) and others have argued that
21 capture of diversity for quantitatively inherited characters requires larger samples (20 – 30% of
22 the whole germplasm collection). The size of the core seems thus to be arbitrary, without a clear
23 objective criterion to judge its adequacy. The kind of species and the extent of the data used in its
24 construction have major influences on the fraction of the whole collection to go into the core.

Besides ecological data, agro-morphological data are another valuable source of information if available for a significant proportion of the collection. Molecular data are probably the most reliable data for core collection development. However, in their absence and also due to the cost of obtaining them for large collections, a combination of ecological and agro-morphological data has been mostly used (Balfourier *et al.* 1999; Jackson *et al.* 1999).

Turning to sampling strategies, many authors have suggested using stratified random sampling strategies to develop core collections (Brown 1989a; Crossa *et al.* 1994; Harch *et al.* 1995). The maximum diversity in the bounds of strata is the main criterion for the selection process (Brown 1995; Diwan *et al.* 1995). Once the percentages of all strata are established, the representatives from these strata will be selected either methodically or randomly. These procedures are based on the size of strata. Other procedures have been suggested when data can be used to estimate the diversity within or between the strata (Schoen and Brown 1993; 1995; Brown and Schoen 1994). The maximising strategy (Schoen and Brown 1993, 1995) focuses on the marker loci with distinct allele types in each accession. This strategy determines the number of accessions that should come from each of the different groups, as well as identification of the accessions that should be included in the core. The expected supremacy of a maximising strategy is due to the link between detected allelic richness at the marker loci and allelic richness on other loci (target loci). These different approaches will be effective if a sound classification with genetically significant groups is carried out (van Hintum *et al.* 2000).

Gouesnard *et al.* (2001) have extended the maximising strategy from marker loci to non-molecular quantitative and qualitative variables. The MSTRAT software (Gouesnard *et al.* 2001) can use different kinds of data to implement a version of this strategy, and can assist in deciding the size of the core collection to be sampled, as well as the choice of the type of genetic richness to be maximized in the core collection. Moreover, the software allows investigation of the

1 amount of retained genetic richness for characters that have not been used in the sampling of the
2 core collection. A general problem with MSTRAT, however, is its absolute focus on minimising
3 or maximising diversity. In other words, the software cannot be used blindly meaning that if the
4 data are strongly heterogeneous, the chance of the most irrelevant accessions being selected in a
5 core due, solely, to having the oddest numbers in the dataset, is very high. This has a very
6 negative impact on the development of a core collection, sometimes resulting in an opposite
7 output to what was originally intended.

8 Genetic resource collections of forage species differ from their crop counterparts in having a
9 much greater role for the provision of genotypes for direct use, rather than just as sources of
10 particular genes for breeders. Although, probe genotypes also exist among the crops but the long
11 domestication history of these genotypes makes them less directly related to environmental
12 factors than forage species, which have been mostly domesticated in a much shorter period of
13 time than crops. It is important therefore that the process of core construction reflects this
14 difference. Both ecological data for the sample site of origin, and agro-morphological data have
15 key roles in designating core entries. Thus procedures are needed that use and integrate these
16 different kinds of data (Prosperi and Sompoux 2005).

17 Annual pasture legumes including species of clover (*Trifolium* L.) are important elements of
18 farming system in southern Australia. Endemic to the Mediterranean region, *Trifolium spumosum*
19 is of particular interest because of its suitability for fine textured soils, high productivity on acid
20 loams and clays both in dry matter and seed yield, and simplicity of seed harvesting using
21 conventional crop machinery (Loi *et al.* 2003). The germplasm collection of this species at the
22 Australian Trifolium Genetic Resource Centre (ATGRC, Perth, Western Australia) comprises
23 389 accessions from 12 countries of origin (Ghamkhar *et al.* 2007). There is a need to establish a
24 robust and reliable core collection of bladder clover for use by breeders for further screening.

There is also a need to expand the core collection methodology to *T. subterraneum*, because of the larger size of its collection and its economic importance in southern Australia. However, the large size (c. 8,000 accessions and genotypes) makes it unsuitable for a pilot study.

In this paper ecogeographical information obtained from the *Trifolium spumosum* passport data at the ATGRC was compared to agro-morphological characteristics for their suitability to develop a core collection. An initial subset was chosen based on the information obtained from the ecological data followed by a final core using both ecological and agro-morphological datasets. The difference between this and other core collections is the combination of maximising and other strategies in a two step development of the core collection. Our aims were to (i) develop a method of core selection of bladder clover that uses ecological information, (ii) investigate the relationships between the morphological and ecogeographical data, (iii) characterise patterns of variation within the species *Trifolium spumosum*, and (iv) compare strategies for choosing a core collection and determine the strategy that best represents the diversity within the bladder clover germplasm collection.

Materials and methods

Collection site data

GIS analysis of the passport and ecogeographic data for the points of origin of 398 accessions of *Trifolium spumosum* conserved in the ATGRC is previously reported (Ghamkhar *et al.* 2007). The key factor used to stratify the collection was the origin of the accessions. Origin is an indicator for both geographical distribution and ecological patterns or environmental regions obtained by the use of GIS. The previous analysis defined the strata within the collection. A total

of 313 accessions from eleven countries had completed ecogeographical and passport data. The countries with numbers of accessions in parenthesis were Algeria (1), Morocco (1), Spain (1), Jordan (2), Iran (3), Italy (4), Cyprus (6), Israel (12), Iraq (13), Syria (38), Greece (88) and Turkey (144). The five ecogeographical variables in the dataset were latitude, longitude, altitude, rainfall, and soil texture. These variables were classified using ANOVA into four categories. Soil texture, however, was classified by combining some of the original seven categories available in the ecological data.

Developing the first subset using ecological data

A step by step process was implemented, ending with approximately 30% of the size of the whole collection as the first (ecological) subset. The first step was choosing the number of accessions to include in the first subset. The second step was an iterative process to capture the highest genetic diversity by selecting 100 samples of 95 accessions using MSTRAT software (INRA 2002). The third step was choosing the final subset of 95 accessions based on their frequency, and the highest Nei and Shannon indices in groups of 95 accessions.

Checking the coverage of the first subset

Ecogeographical criteria

The first subset was tested using the method described by van Hintum *et al.* (2000, p. 28). Means, ranges and variances of ecological and geographic data of the subset were compared with those of the whole collection.

Plant descriptors as refining tools

New agro-morphological characters that played no role in choosing the ecological subset were also used to further test it. In 2004, the 313 bladder clover accessions were evaluated for agro-morphological characters at the Medina Research Station, Western Australia (32.22° S, 115.8° E). A total of six seedlings from each accession was raised in the glasshouse and transplanted into the field five weeks after sowing. The six plants, spaced 40 cm, were grown in a single plot 80 × 120 cm. A total of 10 agro-morphological attributes was scored either visually or quantitatively over five months, between August and December (Table 1). Observations included stem habit and number, stem thickness, leaf size, spring leaf mark, flowering time, calyx pigmentation, seed weight per head, seeds per calyx and 1000 seed weight. Data were subject to an ANOVA. The Mantel test (Mantel 1967) was applied to the complete agro-morphological and ecogeographical datasets to measure their concordance. This was followed by a Canonical Redundancy Analysis (RDA) (Legendre and Legendre 1998) to fix some uncertainties and further clarify the correlation between the agromorphological and ecogeographical datasets using distance matrices of the two. A Principal Component Analysis (PCA) was also performed to summarise the agro-morphological data and assess the importance of the variables. A cluster analysis using UPGMA (Unweighted Pair Group Method with Arithmetic Average) was then carried out on the most important components of the PCA.

Insert Table 1

Further tests using other sampling strategies

Sampling strategies using dendrograms

Ten sampling strategies (Table 2) were employed, all of which randomly generated 95 accessions (about 30% of the whole collection), with the resulting means and variances used to calculate confidence intervals. In the constant strategy, the sample size from each group is equal, independent of the number of accessions in that group (Brown 1989a). In the proportional and logarithmic strategies, entries are allocated to each group in proportion to the number, or to the logarithm of the number of accessions in that group, respectively. Stratification of each ecological cluster was defined in advance by PCA on the original variables. Agro-morphological clusters were created using UPGMA as mentioned above.

Insert Table 2

Testing criteria and comparison of different strategies

Comparisons were made of the means, ranges and variances of ecogeographical and agro-morphological data in the 10 subsets with those of the whole collection. In addition, the core development strategies were tested for their ability to yield a set of samples that produces a distribution map close to the reference map based on the whole collection. Maps of the subsets developed by the ten different sampling strategies were compared with the distribution map of the whole collection. This was done by comparing the most westerly, easterly, northerly, and southerly accessions in the cores with those from the whole collection, as well as representatives from the Mediterranean islands. The preferred strategy is the one with the least difference between the whole and the subset distribution pattern.

Developing the core collection of bladder clover

Three trial cores were finally chosen; one selected randomly, the second using only agro-morphological data, and the third using combined agro-morphological and ecological data (combined core). These selections were developed from the ecological subset, the latter two using MSTRAT. They contained 30% of the number of accessions in the ecological subset making the final size of the core 9% of the whole collection ($30\% \times 30\% = 9\%$). A total of 31 accessions was selected in the final selections using the same steps as in the development of the ecological subset.

Validation of the combined core

A phenetic dendrogram of the whole collection using the agro-morphological data was obtained using NTsys-pc software. This allowed a later comparison of the positions of the members of the combined core on the dendrogram. The Sorenson coefficient (a similarity coefficient, also known as Dice coefficient) (Romesburg 1990) was used prior to UPGMA clustering. UPGMA (contrasting Ward's method) is applicable to morphological and agronomic data. The larger the Sorenson coefficient, the lower is the variance between the two accessions being compared.

The second validation of the combined core was to compare means, ranges, and variances of ecological and agro-morphological data with those of the other selected sets. The final process of validation of the combined core was to compare the same statistical data with those of the whole collection and the ecological subset. A cluster analysis of the 31 accessions in the core was run in NTsys-pc to further validate their geographical distribution.

Results

Maintaining the diversity of the whole collection within the ecological subset

Using MSTRAT, maximum diversity was reached with the sample size of at least 40 accessions (or about 13% of the whole collection) (Figure 1). However the number of accessions in the ecological subset was increased to approximately 30% (95 accessions) to include a larger sample of different combinations of characters.

When tested, the variance for latitude, longitude, altitude, rainfall and soil texture of the collection sites of accessions in the ecological subset exceeded those in the whole collection (Table 3). The maximum values of the five variables in the subset were similar to those in the whole collection (data not shown). Although this signalled an appropriate selection, the accessions with minimum values for longitude and altitude had been omitted. Adding two accessions, one with minimum (the most westerly) longitude (accession number: 44765) and the other with the lowest altitude (accession number: 44771) ensured coverage of the full range.

Testing the ecological subset for representation of agro-morphological diversity supported the subset particularly in terms of variance and range (Table 4).

Principal component and hierarchical cluster analyses

Of 14 agro-morphological traits, five principal components (PCs) for morphological traits and one for an agronomic trait (flowering time) explained 77% of total agro-morphological variation within the whole collection (Table 5). In other words, these six traits vary most among the 14 traits measured here. This also indicates that 23% of the information about the variation will be lost if the other eight traits are not considered for further analyses. The important morphological traits were seeds per calyx, 1000 seed weight, leaf size, calyx pigmentation, and stem thickness. The mean distance (total diversity per trait between paired samples) was 1.33 (relatively low) with a range of 0 – 4.50 (none to very large) for bladder clover accessions. Among the accessions only two from Antalya (Turkey) clustered separately (Figure 2) and may be different taxa.

1

2 *Correlation of the two datasets*

3

4 The Mantel test (Z statistic) between the distance matrices based either on ecogeographical
 5 (ECO) or agro-morphological (MOR) data, gave a matrix correlation value of 0.87, which is

1 considered a good fit. Therefore, it is sensible to combine both ECO and MOR data to develop
2 the final core. However, the resultant graph does not provide a clear visual relationship as it
3 contains far too many points, which prevent the observation of the expected line that reveals the
4 relationship. Further, curvilinear regression analysis revealed almost no relationship between
5 ECO and MOR with negligible R^2 (squared multiple correlation coefficient) values, except the
6 reverse relationship between latitude and leaf size (data not shown). Redundancy Analysis
7 (RDA), however, revealed many unknown correlations with $R^2 = 0.15$ (Figure 3). They included
8 positive correlations of latitude with stem number, rainfall with stem number and flowering time,
9 altitude with flowering time, stem habit, stem thickness, and mean 1000 seed weight; and
10 negative correlations of rainfall with seed weight per head, leaf size, and calyx pigmentation, and
11 altitude with all leaf marks except leaf mark level C.

Figure 3. Ghamkhar et al.

Other selection strategies

Ten other strategies (Table 2) were tested for their ability to retain phenotypic diversity in a core. The strategies SCA, SLA, and SPE gave very similar results to MSTRAT with a slight superiority of SPE for a sample size of 30% of the whole collection (data not shown). The three strategies using combined information, SCC, SPC, and SLC, appeared least efficient overall. By the criteria used here, diversity was not maximised when structures are poorly related. The SLE method also produced a poor result, as did simple random sampling. Two of the three methods using agro-morphological data (SCA and SLA) produced samples with relative diversity indices higher than those of other methods. Clearly, they retained the phenotypic diversity within the

1 whole collection. Overall, agro-morphological data proved to be more informative in retaining
2 the diversity than ecological data.

3
4 *Final Core collection*

5
6 The core developed using combined ecological and agro-morphological data was superior
7 (Tables 3 and 4). The other sampling sets developed either randomly or only with ecological data
8 contained either redundant accessions or did not represent the ecological subset or the whole
9 collection. The final core of 31 accessions using combined data formed a group representing
10 both the whole collection and the ecological subset (Figure 4). It was also well distributed across
11 all agro-morphological and ecogeographical traits except leaf mark level A, rainfall, and altitude
12 (Tables 3 and 4). Accessions 98509 and 34572 were present in all combinations suggesting that
13 they are probably the most accurately selected accessions in the core. Hierarchical cluster
14 analysis supported the selection of the 31 accessions as they grouped into clearly separated
15 clusters (Figure 4).

Figure 4. Ghamkhar et al.

Discussion and conclusion

As noted at the outset, collection of pasture species require the use of ecological and agro-morphological data in the construction of their cores. Ecological data add value to usual passport information and evaluation data that can be directed towards improving genebanks. The aim of this study was to develop a procedure for combining different kinds of data.

The separate use of ecogeographical and agro-morphological traits appears to be the most efficient approach to selecting a first subset before finalising a core. Geographically separate sources of accessions are more likely to have different genetic backgrounds even if they are phenotypically alike. In addition passport data are usually available for most accessions in a collection, whereas trait data are usually much less complete. Therefore our recommended

1 procedure is to use the ecological data for developing a first subset. The strategy for this step is
2 stratified random sampling proportional to group size to yield a total of 30% of the original
3 collection, which, in the bladder clover case, gave a sample size of 95 accessions. This core will
4 be used as 'probe' genotypes, whereby those accessions found to be promising in a particular
5 environment can be used to then identify broadly similar lines in the original collection for more
6 intensive evaluation trials.

7
8 The maximising strategy had no advantage over random strategies at this early stage of
9 developing a core. This strategy chooses a minimum sample size that includes at least one of
10 every variant, or that maximises the variants for a given size. An important factor might explain
11 its poor performance here. The maximising strategy has been designed to include all individual
12 variant classes, without regard to whole genotypes, or eco-geographic origin. It runs the risk of
13 bias towards unusual accessions that possess a combination of odd data. These will be the most
14 diverse and isolated accessions rather than representatives of any group. This is a potential hazard
15 of MSTRAT as if wrongly used, e.g. selecting accessions with too many unique variants; it can
16 mislead the selection process of the core collection.

17
18 The criteria used for the assessment of diversity within the ecological subset are
19 predominantly based on morphological data alone. However, clustering accessions based on
20 morphological traits may be artificial, particularly if the environment under which evaluation
21 took place limited the expression of phenotypes. Such misclassification could fail to accurately
22 represent genetic diversity. Therefore, as was shown here, combining data and analysing them
23 via a maximising strategy was the most efficient technique in the second step of developing the
24 final core collection. This is due to the technique ensuring that most divergent populations or

accessions in the whole collection are included in the core collection. The proposed two step logic in this study and particularly the deterministic approach, SPE, followed by a combined data analysis, is the best approach because it generates optimal results. It will be important to test the core's validity further with molecular markers (Ghamkhar *et al.*, in preparation).

The present study sought evidence as to the core's validity in comparing such ecological and agromorphic group statistics as the mode and the variance of selected versus baseline groups. The modes gave a strong support for high diversity and minimum repetition in the final core. Also, variances of both ecological and agro-morphological traits were consistently higher in the core compared to both the whole collection and the ecological subset, while variances of traits were reduced in the randomly selected core, particularly compared with the whole collection.

Correlation studies of genetic and physical distances are documented for several legumes, with Bessega *et al.* (2000) suggesting no evident correlation while many studies such as Wilkinson *et al.* (1996), Huh and Huh (2001), Dangi *et al.* (2004), and Bushman *et al.* (2007) show remarkable relationships. In this study, RDA showed obvious correlation between ecogeographical conditions and agro-morphological data in bladder clover. In the assessment of the agro-morphological traits, the unimportant traits were mainly the leaf marks, which are probably qualitative and, presumably, represented by a few alleles at few loci, while the important traits were those with quantitative variation. While this may not be unexpected, it is a significant result.

A two step process starting with the SPE strategy followed by using the maximising strategy with the combined ecogeographical and morphological data gave the best results for selecting a core collection. The SCA and SLA procedures also provided acceptable results for the first

1 subset. Combining data for the first subset, however, captured minimum diversity, suggesting
2 that ecogeographical and agro-morphological data should be kept separate in the earlier steps of
3 the procedure. After the first stratification of the whole collection, combining these data and
4 using a maximising strategy will add value to the final choice from the already reduced
5 collection.

6

7 **Acknowledgements**

8

9 Thanks to the Grain Research and Development Corporation (GRDC), Australia for funding this
10 study. Thanks also to the staff of the Australian Trifolium Genetic Resource Centre (ATGRC)
11 who maintain the passport and collection site database. Germplasm collectors who are the
12 originators of this data are also gratefully acknowledged. Special acknowledgment goes to Pierre
13 Legendre, University of Montreal (Canada), for introducing RDA test to the first author. We also
14 thank Doug McEachern, Pro Vice-Chancellor (Research and Innovation) at the University of
15 Western Australia for providing extra funding to the first author for preparing this paper.

References

- Balfourier F, Prosperi JM, Charmet G, Goulard M, Monestiez P (1999) Using spatial patterns of diversity to develop core collections. In 'Core collections for today and tomorrow'. (Eds RC Johnson, T Hodgkin) pp. 37-48. (Internatioanl Plant Genetic Resources Institute: Rome).
- Bessegga C, Ferreyra L, Vilardi JC, Saidman BO (2000) Unexpected low genetic differentiation among allopatric species of section *Algarobia* of *Prosopis* (Leguminosae). *Genetica* **109**, 255-266.
- Bisht IS, Mahajan RK, Lokknathan TR, Agrawal RC (1998) Diversity in Indian sesame collection and stratification of germplasm accessions in different diversity groups. *Genetic Resources and Crop Evolution* **45**, 325-335.
- Brown AHD (1989a) Core collections: a practical approach to genetic resources management. *Genome* **31**, 818-824.
- Brown AHD (1989b) The case for core collections. In 'The use of plant genetic resources'. (Eds AHD Brown, OH Frankel, DR Marshall, JT Williams) pp. 136-156. (Cambridge University Press: Cambridge, UK).
- Brown AHD (1995) The core collection at the crossroads. In 'Core collections of plant genetic resources'. (Eds T Hodgkin, AHD Brown, TJL van Hintum, EAV Morales) pp. 3-20. (John Wiley and Sons: New York).

1 Brown AHD, Schoen DJ (1994) Optimal sampling strategies for core collections of plant genetic
2 resources. In 'Conservation genetics'. (Eds V Loeschckje, T Omiuk, SK Jain) pp. 357-370.
3 (Birkhauser Verlag: Basel, Switzerland).

4
5 Bushman BS, Larson SR, Peela MD, Pfrender ME (2007) Population structure and genetic
6 diversity in North American *Hedysarum boreale* Nutt. *Crop Science* **47**, 1281-1288.

7
8 Charmet G, Balfourier F (1995) The use of geostatistics for sampling a core collection of
9 perennial ryegrass populations. *Genetic Resources and Crop Evolution* **42**, 303-309.

10
11 Crossa J, Taba S, A. ES, P. B, R. V (1994) Practical considerations for maintaining germplasm in
12 maize. *Theoretical and Applied Genetics* **89**, 89-95.

13
14 Dangi R, Lagu M, Choudhary L, Ranjekar P, Gupta V (2004) Assessment of genetic diversity in
15 *Trigonella foenum-graecum* and *Trigonella caerulea* using ISSR and RAPD markers. *BMC*
16 *Plant Biology* **4**, 13.

17
18 Diwan N, McIntosh MS, Baughan GR (1995) Methods of developing a core collection of annual
19 *Medicago* species. *Theoretical and Applied Genetics* **90**, 775-761.

20
21 Ghamkhar K, Snowball R, Bennett SJ (2007) Ecogeographical studies identify diversity and
22 potential gaps in the largest germplasm collection of bladder clover (*Trifolium spumosum* L.).
23 *Australian Journal of Agricultural Research* **58**, 728-738.

- 1 Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: An
2 algorithm for building germplasm core collections by maximizing allelic or phenotypic
3 richness. *Journal of Heredity* **92**, 93-94.
4
- 5 Harch BD, Brasford KE, DeLacy IH, Lawrence PK, Cruickshank A (1995) Patterns of diversity
6 in fatty acid composition in the Australian groundnut germplasm collection. *Genetic
7 Resources and Crop Evolution* **42**, 243-256.
8
- 9 Huh MK, Huh HW (2001) Genetic Diversity and Population Structure of Wild Lentil Tare. *Crop
10 Science* **41**, 1940-1946.
11
- 12 INRA (2002) 'MStrat documentation.' (www1.montpellier.inra.fr/gap/MSTRAT/doc/index.html)
13 Accessed May. 2008.
14
- 15 Jackson MT, Pham JL, Newbury HJ, Ford-Lloyd BV, S. VP (1999) A core collection for rice -
16 needs, opportunities, and constraints. In 'Core collections for today and tomorrow.'. (Eds RC
17 Johnson, T Hodgkin) pp. 18-27. (IPGRI: Rome, Italy).
18
- 19 Legendre P, Legendre L (1998) 'Numerical ecology.' (Elsevier Science BV: Amsterdam).
20
- 21 Loi A, Nutt BJ, McRobb R, Carr SJ (2003) *Trifolium spumosum* L. an exciting prospective
22 legume for fine textured soils in Mediterranean farming systems. In 'The proceedings of the
23 11th Agronomy Conference'. (Eds M Unkovich, G O'Leary) pp. 1-4. (The Australian Society
24 of Agronomy: Geelong, Victoria, Australia).

1

2 Mantel N (1967) The detection of disease clustering and a generalized regression approach.

3 *Cancer Research* **27**, 209-220.

4

5 Noirot M, Hamon S, Anthony F (1996) The principal component scoring: a new method of

6 constituting a core collection using quantitative data. *Genetic Resources and Crop Evolution*7 **43**, 1-6.

8

9 Prosperi JM, Sampoux JP (2005) Conservation of the genetic resources of forage plants: what is

10 the desired level of genetic diversity within a collection? *Fourrages* **182**, 245-262.

11

12 Romesburg HC (1990) 'Cluster analysis for researchers.' (Krieger Publishing: Malabar, FL,

13 USA).

14

15 Schoen DJ, Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by

16 assessment of genetic markers. *Proceeding of the National Academy of Sciences. USA* **90**,

17 10623-10627.

18

19 Schoen DJ, Brown AHD (1995) Maximising allelic diversity in core collections of wild crop

20 relatives: The role of genetic markers. In 'Core Collections of Plant Genetic Resources'. (Eds T

21 Hodgkin, AHD Brown, TJL van Hintum, EAV Morales) pp. 55-79. (John Wiley and Sons:

22 New York).

23

1 van Hintum TJL, Brown AHD, Spillane C, Hodgkin T (2000) Core collections of plant genetic
2 resources. The International Plant Genetic Resources Institute (IPGRI) Bulletin No. 3, Rome,
3 Italy.

4

5 Wilkinson HH, Spoerke JM, Parker MA (1996) Divergence in symbiotic compatibility in a
6 legume-Bradyrhizobium mutualism. *Evolution* **50**, 1470-1477.

7

8 Yonezawa K, Nomura T, Morishima M (1995) Sampling strategies for use in stratified
9 germplasm collections. In 'Core Collections of Plant Genetic Resources'. (Eds T Hodgkin,
10 AHD Brown, TJL van Hintum, EAV Morales) pp. 35-53. (John Wiley and Sons: Chichester,
11 UK).

12

13

Figure captions

Figure 1. Diversity score created by MSTRAT for different sample sizes of the first subset of accessions using maximising (M method) and random sampling strategies.

Figure 2. Hierarchical cluster analysis of the agro-morphological data in the original collection of *T. spumosum* using sequential agglomerative hierarchical and nested (SAHN) clustering and the origin of each cluster. Accessions originating from the same regions are identified using a variety of symbols which are reflected on the map above.

Figure 3. Redundancy Analysis (RDA) biplot showing correlation between ecogeographical and agro-morphological traits. RDA 1 and RDA 2 represent principal components 1 and 2, respectively, obtained after regression analysis. Solid arrows represent the agro-morphological variables and dashed arrows represent ecogeographical (site) variables. Codes of agro-morphological traits are listed in Table 1.

Figure 4. Hierarchical cluster analysis of the agro-morphological data of the 31 accessions of the core collection using SAHN clustering. Accessions are listed on the right with corresponding geographic regions and countries of origin.

Table 1. Number of replicates per plant assessed and number of states adopted for a range of agro-morphological traits, in *T. spumosum* (description of states for each trait in parenthesis).

Agro-morphological trait	No. of replicates per plant	No. Of states per trait	Trait code
Seed number per calyx	4		SPC
1000 seed weight (g)	2		MSW
Leaf size (1 = small, 10 = large)	1	10	LS
Calyx pigmentation (1 = absent, 6 = all red)	1	6	CP
Stem thickness (mm)	2		ST
Flowering time (days)	1		FT
Stem number (5, 17, 33, 70, 150)	1	5	SN
Stem habit (1 = prostrate, 3 = intermediate, 5 = upright)	1	5	SH
Seed weight per head (g)	1		SWPH
Spring leaf mark level A (crescent, arm, crescent and arm, no mark, miscellaneous)	1	5	LMLA
Spring leaf mark level B (no crescent flush and wings, crescent flush or wings, no red arm, red arm)	1	4	LMLB
Spring leaf mark level C (general flush, general flush and wings, wings, mid-rib flush)	1	4	LMLC
Spring leaf mark level D (average flush, strong flush)	1	2	LMLD
Spring leaf mark level E (average red arms, weak red arms)	1	2	LMLE

Table 2. Ten sampling strategies used in this study to validate the first subset.

Stratification	Sampling method	Data type	Code
Not stratified	Simple		S
Stratified	Constant	Ecogeographic	SCE
		Agro-morphological	SCA
		Combined	SCC
	Proportional	Ecogeographic	SPE
		Agro-morphological	SPA
		Combined	SPC
	Logarithmic	Ecogeographic	SLE
		Agro-morphological	SLA
		Combined	SLC

Table 3. Statistical data of the ecogeographic variables for the whole collection, first (ecological) subset, random core and final (combined data) core collection.

	Latitude	Longitude	Altitude (m)	Rainfall (mm)	Soil texture [*]
Whole collection					
Variance	2.5	51.31	112588	40450.4	2
Mode	36	25.43	500	500	2
Median	37.21	33.83	283.5	600	2
Range	9.3	53.54	1660	1275	6
Mean	37.08	32.34	368.6	628.1	2.4
Mean-median	0.12	1.49	85.1	28.1	0.4
First subset					
Variance	2.5	55.16	123379.2	44901.7	2.3
Mode	37.55	25.16	500	500	2
Median	37	34.75	250	600	2
Range	8.43	48.60	1460	1250	6
Mean	36.72	32.52	362.7	628.4	2.4
Mean-median	0.28	2.22	112.7	28.4	0.4
Random core					
Variance	2.04	42.36	96090.5	18667.4	14.1
Mode	N/A	N/A	800	500	10
Median	37.21	36.82	487.5	512.5	10
Range	6.94	23.71	990	650	9
Mean	37.23	34.92	446.7	554.2	6.7
Mean-median	0.02	1.90	40.8	41.7	3.3
Final core					
Variance	2.5	114.84	172508	48604.5	2.8
Mode	37.47	N/A	800	400	1
Median	37.16	33.85	335	500	2.5
Range	7.47	48.60	1450	750	5
Mean	36.95	31.34	456.1	567.9	2.7
Mean-median	0.21	2.50	121.1	67.9	0.2

^{*} quantified in a 1-7 scale: 1: loam, 2: sandy loam, 3: clay loam, 4: sand, 5: clay, 6: gravel, 7: stones.

Table 4. Statistical data of agro-morphological variables for the whole collection, first (ecological) subset, the random core, and the final (combined data) core collections.

	SPC	MSW	LS	CP	ST	FT	SN	SH	SWPH	LMLA	LMLB	LMLC	LMLD	LMLE
Whole collection														
Variance	0.31	0.22	1.47	0.42	1.55	43.29	174.60	0.04	0.01	0.23	0.8	0.01	0.008	0.30
Mode	4	2.51	2.50	3	5.40	139	45	1.30	0.50	1	1	0	0	0
Median	3.79	2.41	2.83	3	6.10	133.67	42.50	1.30	0.39	1	1	0	0	0
Range	3.75	2.98	7.67	4.5	7.80	50.83	70.60	1	0.84	4	4	1.50	0.80	2
Mean	3.60	2.41	3.08	3.15	6.18	133	43.82	1.29	0.39	1.15	1.35	0.02	0.01	0.20
Mean-median	0.19	0	0.25	0.15	0.08	0.67	1.32	0.01	0	0.15	0.35	0.02	0.01	0.20
First subset														
Variance	0.28	0.20	1.35	0.39	1.39	34.55	134.61	0.04	0.01	0.04	0.37	0.02	0.005	0.16
Mode	3.95	2.67	2.50	3	5.90	139	51.30	1.30	0.40	1	1	0	0	0
Median	3.75	2.50	3	3	6.10	133.50	42.50	1.30	0.38	1	1	0	0	0
Range	3.71	2.53	6.80	4.50	7.10	38.66	70.20	1	0.74	1	4	1.50	0.80	2
Mean	3.59	2.47	3.13	3.10	6.20	132.96	42.69	1.30	0.39	1.05	1.19	0.03	0.01	0.1
Mean-median	0.16	0.03	0.13	0.10	0.10	0.54	0.19	0	0.01	0.05	0.19	0.03	0.01	0.1
Random core														
Variance	0.15	0.11	0.90	0.34	1.58	35.25	218.06	0.07	0.005	0.09	0.71	0.02	0.004	0.25
Mode	3.96	2.57	2.50	3	6.70	130	63.80	1.20	0.40	1	1	0	0	0
Median	3.94	2.57	2.67	3	6.35	134.58	45	1.30	0.38	1	1	0	0	0
Range	1.91	1.38	4.16	2	5.30	23	66	1	0.33	1	3	0.70	0.30	2
Mean	3.79	2.58	2.95	3.26	6.37	133.92	46.90	1.35	0.38	1.11	1.34	0.03	0.02	0.16
Mean-median	0.14	0.01	0.28	0.26	0.02	0.66	1.90	0.05	0	0.11	0.34	0.03	0.02	0.16
Final core														
Variance	0.77	0.27	2.14	0.58	2.40	63.18	286.30	0.06	0.02	1.25	1.20	0.14	0.05	0.38
Mode	4.04	2.35	2.67	3	5.40	139	57.50	1.30	0.60	1	1	0	0	0
Median	3.63	2.33	2.67	3	6.10	135	40	1.30	0.38	1	1	0	0	0
Range	3.38	2.37	6.80	4	7.10	38.66	70.20	1	0.64	1	4	1.50	0.80	2
Mean	3.37	2.31	3.16	3.14	6.08	133.29	44.86	1.31	0.39	1.48	1.43	0.18	0.09	0.23
Mean-median	0.25	0.02	0.49	0.14	0.02	1.71	4.86	0.01	0.01	0.48	0.43	0.18	0.09	0.23

Table 5., Eigenvalues and percentage and cumulative percentage of total diversity of agro-morphological traits within the whole collection. Traits in bold account for 76.8 percent of diversity of *T. spumosum*.

Trait	Eigenvalue	Proportion	Cumulative
Mean Number of seeds per calyx	2.8	19.7	19.7
Mean 1000 seed weight	2.6	18.4	38.1
Leaf size	1.7	12.4	50.5
Calyx pigmentation	1.4	10.4	60.9
Stem thickness	1.2	8.5	69.4
Flowering time	1.0	7.4	76.8
LMLA	0.9	6.2	83.0
Stem number	0.6	4.2	87.2
Stem habit	0.5	3.8	91.0
LMLB	0.4	2.8	93.8
LMLC	0.3	2.5	96.3
LMLD	0.3	2.4	98.7
LMLE	0.1	1.1	99.8
Seed weight per head	0.025	0.182	~ 100%